Technical Report 840

A Review of Procedures for Setting Job Performance Standards

Elaine Pulakos
Human Resources Research Organization

Lauress Wise
American Institutes for Research

Jane Arabian
U.S. Army Research Institute

Susan Heon and S. Kathleen Delaplane
Human Resources Research Organization

May 1989

United States Army Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

JON W. BLADES
COL, IN
Commanding

Research accomplished under contract for the Department of the Army

American Institutes for Research

Technical review by

Frances C. Grafton
Michael G. Rumsey

NOTICES

DISTRIBUTION : Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POX, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600

FINAL DISPOSITION : This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE : The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS --
2a. SECURITY CLASSIFICATION AUTHORITY --	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE --	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) --	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Technical Report 840
6a. NAME OF PERFORMING ORGANIZATION American Institutes for Research 6b. OFFICE SYMBOL (If applicable) WRC	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences
6c. ADDRESS (City, State, and ZIP Code) 3333 K Street, NW Washington, DC 20007	7b. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333
8a. NAME OF FUNDING/SPONSORING ORGANIZATION -- 8b. OFFICE SYMBOL (If applicable) --	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA903-87C-0525
8c. ADDRESS (City, State, and ZIP Code) --	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO.: 62722A; PROJECT NO.: 791; TASK NO.: 231; WORK UNIT ACCESSION NO.: C1

11. TITLE (Include Security Classification)
A Review of Procedures for Setting Job Performance Standards

12. PERSONAL AUTHOR(S) Pulakos, Elaine (HumRRO); Wise, Lauress (AIR); Arabian, Jane (ARI); Heon, Susan, and Delaplane, S. Kathleen (HumRRO)

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM 87/04 TO 89/05	1989, May	59

16. SUPPLEMENTARY NOTATION
Jane M. Arabian, Contracting Officer's Representative, Selection and Classification Technical Area.

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Standard setting Job performance Selection

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This report summarizes recent literature on setting job performance standards. Major topics include (a) Army uses of job performance standards, (b) alternative judgment paradigms, (c) factors affecting the judgment process, (d) combining multiple standards, and (e) linking performance standards to selection test scores. A model of the standard setting process is also included. The report concludes with recommendations for further research.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Jane M. Arabian	22b. TELEPHONE (Include Area Code) (703) 274-8275 22c. OFFICE SYMBOL

DD Form 1473, JUN 86 *Previous editions are obsolete.*

Technical Report 840

A Review of Procedures for Setting Job Performance Standards

Elaine Pulakos
Human Resources Research Organization

Lauress Wise
American Institutes for Research

Jane Arabian
U.S. Army Research Institute

Susan Heon and S. Kathleen Delaplane
Human Resources Research Organization

Selection and Classification Technical Area
Frances Grafton, Acting Chief

Manpower and Personnel Research Laboratory
Curtis Gilroy, Acting Director

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

May 1989

Army Project Number
2Q162722A791

Synthetic Validation

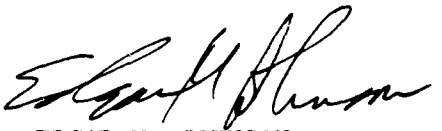
FOREWORD

In 1980 the Assistant Secretary of Defense directed all services to validate the Armed Services Vocational Aptitude Battery (ASVAB) and to re-evaluate enlistment standards against on-the-job performance. The Army has been investigating the validity of the ASVAB and several new predictor measures for a sample of 20 diverse MOS. This effort, known as Project A, has been very successful in validating the ASVAB and in providing the Army with a greater understanding of the Knowledge, Skills, Abilities, and Other personal characteristics (KSAOs) required for these 20 MOS.

A major question now facing the Army is how to extend the wealth of data collected for Project A to the other 25-plus entry-level Army MOS and to new MOS created for new hardware systems as they become operational. A second challenge is to determine the best methods for setting job performance standards that can be used in making selection and classification decisions.

The Synthetic Validation Project (SYNVAL) addresses these challenges. Specifically, the objectives of SYNVAL are to: (1) Evaluate synthetic validation techniques for determining MOS-specific selection composites for each MOS; and (2) evaluate alternative methods for setting minimum qualifying scores on each of these composites. The research will proceed in three phases.

Based on the results of the evaluations, recommendations will be made for the following: (1) A methodology for developing job performance prediction equations for all of the Army's 250-plus MOS and (2) a methodology for setting performance standards for these MOS. The technical quality of this project is guided by the Scientific Advisory Committee, Drs. Phil Bobko (Chair), Robert Linn, Richard Jaeger, Joyce Shields, and Robert Guion.



EDGAR M. JOHNSON
Technical Director

A REVIEW OF PROCEDURES FOR SETTING JOB PERFORMANCE STANDARDS

EXECUTIVE SUMMARY

Requirement:

The overall goal of the project is to develop a system for choosing valid selection and classification tests for all Army MOS and for setting minimum selection standards on these tests. Initially, a literature review was conducted to identify factors contributing to the quality and acceptance of standard setting methodologies.

Procedure:

Researchers initiated a computer search to locate relevant literature published in the last 5 years and formulated a model that took into account the variables identified in the research.

Findings:

Three general types of standard setting procedures were reviewed: item-based methods, examinee-based methods, and outcome-based methods. This review indicates that the quality and acceptance of the standards that are developed depend on an interaction among the procedures employed, the characteristics and training of the judges, the types of measures for which standards are to be set, and the overall purpose for setting standards. The beginning of this report proposes a model describing the major interactions among these different components.

The report discusses several aspects of the judgment process, including the number and characteristics of the judges, the training provided to the judges, and the use of judgment facilitation techniques. Methods for combining multiple performance standards into an overall standard (e.g., compensatory and multiple hurdle models) are reviewed, and procedures for linking job performance standards to selection test standards are addressed.

Utilization of Findings:

These results will be used to develop instruments that describe performance levels. These instruments include: (a) the Soldier-based instrument, in which officers and NCOs will be asked to estimate the percentage of soldiers they have seen performing Unacceptably, Marginally, Acceptably, and in an Outstanding manner (UMAO); (b) the Critical Incident instrument, in which officers and NCOs will be asked to rate critical incidents as UMAO; and

(c) the Task-based instrument, in which respondents will be asked to rate various levels of performance as UMAO. These three approaches for setting performance standards will be evaluated in field tests with Army subject matter experts.

A REVIEW OF PROCEDURES FOR SETTING JOB PERFORMANCE STANDARDS

CONTENTS

LIST OF FIGURES

A REVIEW OF PROCEDURES FOR SETTING JOB PERFORMANCE STANDARDS INTRODUCTION

The word standard has a wide range of meanings. Popham (1978) discusses some of the more common definitions of a standard and selects the following from the Oxford English Dictionary as most relevant to performance standards:

> "A definite level of attainment, wealth, or the like, or a definite degree of any quality viewed as a prescribed object of endeavor or as a measure of what is adequate for some purpose."

A key implication of this definition is that standards do not exist in the absolute but rather are necessarily related to some particular purpose.

There has been considerable research focused on developing and evaluating procedures for setting standards, in areas ranging from educational testing to professional licensure. Along with numerous empirical studies, there also have been a number of recent, reasonably comprehensive reviews of the standard setting literature. One strikingly apparent conclusion from this literature is that standard setting has been almost exclusively concerned with establishing minimum scores for passing tests, especially unidimensional, multiple choice tests. Virtually no research has attempted to apply standard setting methodologies to other types of measures, such as job performance measures. The purpose of the present paper is to review the standard setting literature with a specific focus on implications for setting job performance standards. Of particular interest here is the development of procedures for setting performance standards in Army enlisted jobs or Military Occupational Specialties (MOS). Potential uses of job performance standards are much broader than those of certification standards. Job performance standards can provide useful input in several personnel activities, including (1) employee motivation, (2) identification of training needs, (3) evaluation of personnel programs, and (4) setting minimum entry standards.

With respect to motivating employees, the goal setting literature suggests that motivation will be maximized by goals that are specific, challenging, and accepted (Locke, Shaw, Saari & Latham, 1981). Since performance standards provide specific

targets for individual employees, these standards could be used as a basis for setting goals and motivating individuals to higher levels of achievement. For example, marginal employees could be encouraged to meet minimum performance standards, while employees interested in advancement could be encouraged to meet performance standards used to determine eligibility for promotion.

Job performance standards can also be used to identify training needs. Employees who are not performing up to minimum standards may be directed into remedial training courses. In some instances, when an employee is performing far below standards, the organization may decide it would be better to terminate the employee instead of providing him or her with remedial training. It may thus be useful to consider setting multiple performance standards for different personnel decisions. These standards could be used to determine when to terminate employees for poor performance, provide employees with remedial training in response to marginal performance, or reward employees for excellence.

A third potential use for job performance standards is as a yardstick for measuring the effectiveness of entire personnel programs or interventions. If a new personnel system is operating effectively, it should increase the number of employees who are performing satisfactorily (i.e., at or above minimum standards). If a significant number of employees fail to perform up to standards, it may suggest that changes in selection, training, or other personnel programs are warranted.

Lastly, job performance standards are useful for setting minimum entry requirements. By identifying minimal levels of acceptable performance for various jobs, one can determine the selection test score levels that lead to a reasonable probability that the performance standards will be met.

While job performance standards could prove quite useful to organizations, there are several challenges associated with their development and use. First, the vast majority of standard setting methods are designed to yield one minimum "cutting score" (or standard) for passing a multiple choice test. As mentioned above, however, it frequently may be desirable to develop multiple job performance standards for different personnel decisions. Further complicating the process of setting performance standards is that job performance is multidimensional (e.g., requires technical knowledge,

leadership skills), with multiple possible measures of each dimension. Again, this differs from most previous standard setting research that has dealt almost exclusively with unidimensional measures. Finally, the measures most often used to evaluate job performance are not, nor should they necessarily be, multiple choice test items. Multiple choice knowledge tests can be thought of as "maximal" performance measures, while other measures such as ratings provide a better indication of an employee's "typical" job performance over time. In addition, knowledge is a necessary but not sufficient requirement for effective job performance. It is possible, for example, that an individual can have the requisite knowledge of how to perform a task without actually being able to perform that task effectively. Thus, use of measures such as ratings or hands-on test scores require an expanded consideration of relevant methodology in attacking the more general problem of setting performance standards.

In summary, our purpose here is to review the standard setting literature with a specific focus on implications for setting job performance standards, in general, and Army performance standards, in particular. However, given that there are several unique challenges associated with the development of job performance standards that have not been addressed in previous standard setting research, additional literature relevant to these issues will also be reviewed. In the next section, we begin by outlining the particular needs of the Army and the intended uses of job performance standards in that context.

SPECIFIC ARMY USES OF JOB PERFORMANCE STANDARDS

Background

The Army's primary interest in setting job performance standards is so that these standards can be used as a basis for setting enlistment (i.e., selection) standards. Linkage of applicant test scores to subsequent job performance levels will inform decisions regarding selection test screening scores and will help to set targets for the number and distribution of "high quality" accessions. In addition, as new jobs are created, the Army will have the ability to forecast performance requirements and, in turn, identify appropriate applicant pools.

Currently, the Army uses written tests, Skill Qualification Tests (SQT), to assess job knowledge of enlisted soldiers at various skill levels. These are job-specific knowledge tests that measure a soldier's knowledge of how to perform specific tasks required in his or her MOS. While the SQT has been effective in identifying training deficiencies, their use in operational personnel decisions has been somewhat limited in that the SQT is only one of several factors used in determining eligibility for advancement. In part, this is due to the fact that different tests have varying psychometric qualities and that the equating of a test for a given job from one year to the next is relatively imprecise. Although SQT scores have been used in validating selection methods (Hanser & Grafton, 1983; McLaughlin, Rossmeissl, Wise, Brandt, & Wang, 1984), SQT standards are only just beginning to be used in setting selection test cutoffs using a version of the contrasting groups method described below (TRADOC Reg 351-2).

Project A and the Army Synthetic Validation Project

Recently, the Army and the other Services have undertaken Job Performance Measurement (JPM) research projects to determine whether job performance could be measured reliably and if so, how job performance data could be used in setting enlistment standards (see Wigdor & Green, 1986 for a complete discussion of this research). The Army's effort, known as Project A, is a seven-year longitudinal validation of current and alternative selection tests against a wide array of job performance measures.

The Project A research is being conducted on a sample of 21 MOS, carefully selected to be representative of the entire population of Army MOS. Ultimately, however, the Army must develop selection measures and set minimum scores for more than 250 entry-level MOS. Further, new MOS are continually being created and selection procedures will be needed for these jobs as well. Thus, a second large-scale research effort, called The Army Synthetic Validation Project, was undertaken. The purposes of this project are:

- To evaluate the application of synthetic validation procedures in identifying appropriate composites of selection tests for Army enlisted MOS; and,

- To develop procedures for setting selection test standards that are linked to standards for job performance (Wise, Campbell, & Arabian, 1987).

The Project A predictor and criterion measures will be used in developing procedures to set Army performance standards and to link these standards to enlistment standards. Several job performance measures for evaluating first and second tour soldier effectiveness have resulted from Project A. The first tour performance measures include:

- Hands-on tests for each of a carefully selected sample of 15 tasks, with each task scored in terms of the percentage of the steps the examinee performed correctly;

- Job knowledge tests consisting of multiple choice items, that measure knowledge necessary to perform each of 30 carefully selected tasks (including the 15 tasks tested in the hands-on mode);

- Supervisor and peer ratings of 11 common dimensions of performance (e.g., technical proficiency, leadership, integrity) and 7-12 job specific aspects of performance; and,

- Performance indicators derived from administrative records (e.g., awards and certificates, disciplinary problems, physical readiness scores).

Beyond validating the selection instruments against first tour soldier performance, the Project A research also involves validating these measures against second tour effectiveness, for those soldiers who reenlist in the Army. One performance component that is unique to the second tour (versus first tour) soldier's job is supervision. In particular, there are several supervisory behaviors (e.g., counseling subordinates, training subordinates) that are required of all Non-Commissioned Officers (NCOs), regardless of their MOS. In addition to developing second tour technical proficiency measures like the first tour measures listed above, measures of supervisory effectiveness were also developed to evaluate second tour performance. It should thus be possible to set performance standards for both first and second tour positions. In turn, linking selection test scores to subsequent higher-level performance standards

should allow for a more refined definition of "quality" recruits for each career field and could lead to an improved system for assuring future leaders.

Specific Army Needs

Determining minimum qualification levels for enlisted jobs is only part of what is required to assure an adequate combat force. In addition to having soldiers who perform their jobs adequately (i.e., meet the minimum job requirements), some soldiers whose abilities and performance surpass the job requirements are also needed to form the basis of a competent NCO corps. Individuals with lower aptitudes and abilities may satisfy entry-level job performance standards, but promotion to higher level jobs, requires the demonstration of advanced skills such as leadership. Since the Army fills all higher level positions through promoting the potential for developing these advanced skills, they must be considered at the time of initial selection. Indeed, the Army recognizes the need for leaders within each occupation by setting explicit goals for the proportion of quality recruits to be selected into each MOS.

An adequate combat force thus requires an appropriate distribution of quality soldiers. That is, the combat force should be composed of some percentage of at least minimally competent soldiers and some percentage of higher ability soldiers. Statistics such as the percent performing above some minimum standard or even the mean performance level do not convey any information on the degree of variation in performance levels. At least two separate standards would be necessary to convey information on variability. Within Army jobs, for example, it might be appropriate to establish a standard for excellence, indicating future leadership potential, and a minimum standard, reflecting satisfactory performance within current grade.

For the Army's purposes, the standard setting process will require developing procedures that:

- can be applied to performance measures of the type developed in Project A;
- yield reliable, multiple performance standards;
- indicate how standards reflecting multiple dimensions of performance should be combined into an overall standard; and,

- provide a mechanism for linking performance standards to enlistment standards.

Although all four of these issues will be addressed in this paper, we will first review existing standard setting methods and variables that have been shown to influence the standard setting judgment process. The next section presents a model that is used as a framework for organizing this discussion.

A MODEL OF THE STANDARD SETTING PROCESS

A number of factors have been shown to influence the standard setting process and consequently, the properties and acceptance of the resultant standards. Properties of a standard include both its level and the degree of consensus among standard setting judges about that level. Factors affecting the standard setting process include:

- The purpose(s) for setting the standards;
- The types of measures for which the standards must be set;
- The procedures selected for setting standards, including the judgment paradigm itself as well as any judgment facilitation techniques (e.g., employing an iterative judgment process);
- Personal characteristics of the judges;
- The training provided to judges; and,
- The number of judges used to set standards.

To organize this material, the model in Figure 1 was developed. Although this model represents our interpretation of the major variables and relationships addressed by the standard setting literature, it is intended more as a mechanism for organizing the first part of this review than as a comprehensive theory of the standard setting process.

For the present purposes, Standard Setting Procedures refer both to the judgment paradigm (i.e., standard setting method) selected for setting standards and to the use of any judgment facilitation techniques (e.g., providing normative data to judges, employing an iterative judgment process). However, we then proceed to discuss judgment facilitation techniques in a section separate from that in which the judgment paradigms themselves are presented. This is because the judgment facilitation techniques discussed here can be used in conjunction with virtually any standard setting method.

The model of standard setting shown in Figure 1 is composed of eight major categories of variables. As shown in the figure, we propose that the choice of Standard Setting Procedures will be a function of two primary factors. These are:

- the Purposes for setting standards, including the intended use of the standard(s) and the number of standards required (path a in Figure 1) and

- the Types of Measures (path b) for which standards must be set (e.g., multiple choice tests, performance ratings, structured interviews, etc.).

The Standard Setting Procedures, in turn, have a direct influence on three variables. First, it will be argued that the procedures used dictate the type of Judge Training that should be provided (path c in Figure 1). Second, the Standard Setting Procedures influence the Properties of the Standard(s) obtained (path d). In particular, different procedures can lead to different levels of consensus among judges as well as different absolute standards (e.g., some procedures have been shown to result in much more stringent standards than others). Finally, the Standard Setting Procedures dictate, at least in part, the Characteristics judges should possess (path e). That is, different procedures require that judges have different types of knowledge or expertise. For example, some procedures require judges to be familiar with particular examinees' performance while others require judges to be knowledgeable about test items.

The Properties of the Standard(s) will also be a direct function of: 1) Judge Training procedures (path f), 2) Judge Characteristics (path g), and 3) the Number of Judges used to set standards (path h). First, we propose that providing appropriate training to judges can lead to higher quality standards, both in terms of consensus and

absolute level. With respect to judge characteristics, demographic variables (e.g., race) can affect judges' recommended standards. In addition, there may be less consensus achieved and even different absolute standards recommended when judges represent different constituencies or interest groups. Third, the number of judges required to achieve a quality standard is that number which yields an acceptable standard error of the recommended standard. However, this standard error depends not only on the number of judges used but also on the consensus among these judges. Several variables in the model can indirectly influence the number of judges required through their direct affect on consensus. For example, if particular judge characteristics and/or training procedures serve to increase consensus among judges, fewer judges may be required.

The arrow from Judge Characteristics to Number of Judges (path i in Figure 1) was included because we will argue that statistical considerations may not be the only important determinant of how many judges should be used. Other factors, such as ensuring that all relevant constituencies are represented in the group of standard setting judges, should also be considered in determining final sample sizes.

Finally, the model assumes that two variables will directly affect Acceptance of Standards. First, of course, is the Properties of the Standard(s) obtained (path j). Second, related to the point made above, is that acceptance of the standard is likely to be greatly facilitated if an attempt is made to involve relevant constituencies or interested parties in the entire standard setting process. Thus, Judge Characteristics are likely to have a direct effect on standard acceptance (path k).

The next two major sections of this review focus on research relevant to our model. In particular, studies investigating standard setting methods and factors influencing the standard setting judgment process are reviewed. As mentioned previously, the intent here is not simply to review this literature but also to examine its implications for setting job performance standards, in general, and Army performance standards, in particular. Following this review, possible approaches for dealing with the multidimensionality of job performance will be discussed. Of primary concern here is how to best combine multiple standards reflecting different performance dimensions. The final section of this paper will then focus on issues involved in linking selection procedures to performance standards. One important topic

here is evaluating the tradeoffs between selecting individuals who then do not perform up to standards (i.e., false positives) and rejecting individuals who would have performed up to or surpassed the standards (i.e., false negatives).

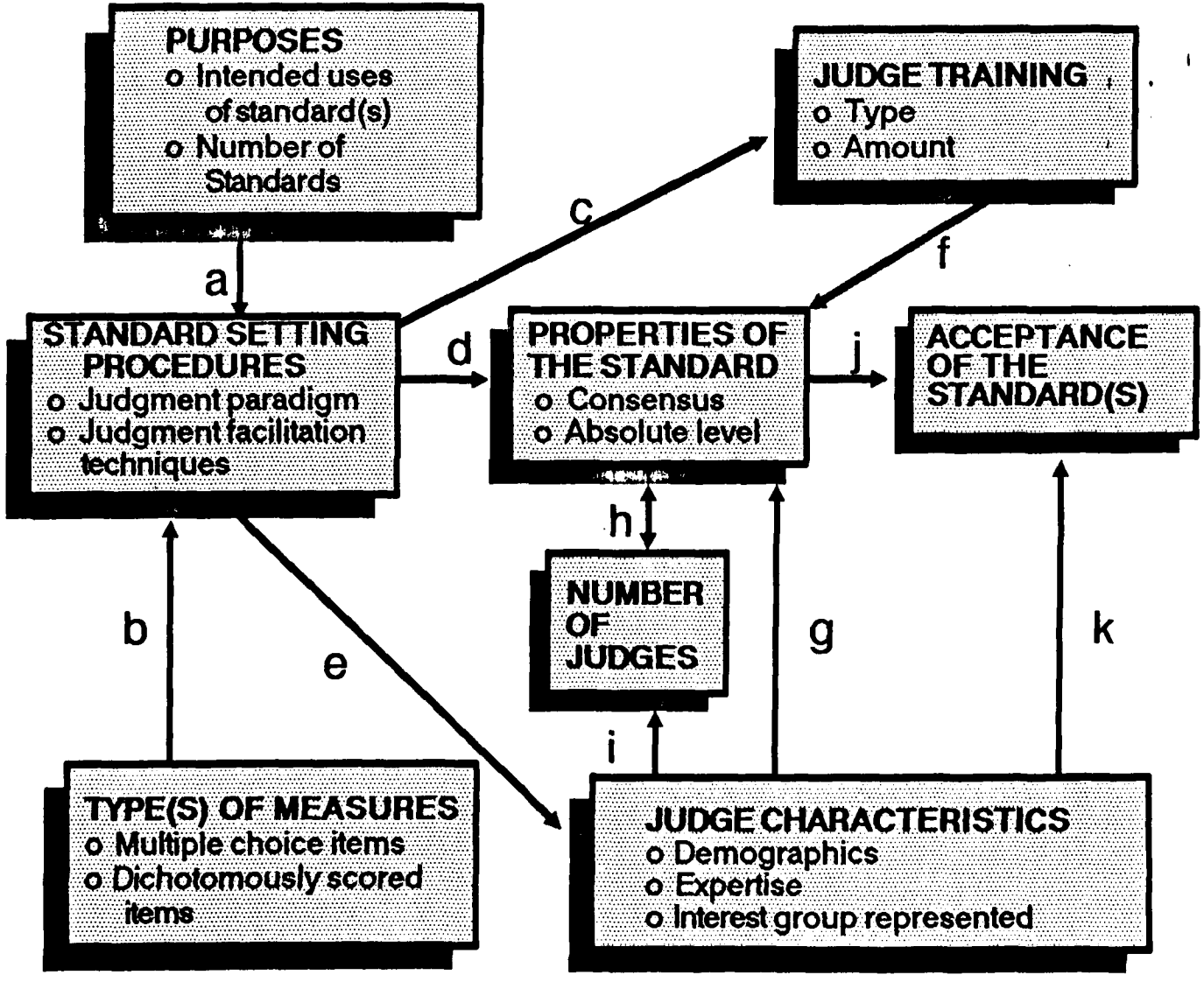


Figure 1. A proposed model of the standard setting process.

JUDGMENT PARADIGMS

Prior reviews (e.g., Berk, 1986; Arabian, 1986; Shepard, 1984; Hambleton, 1980; Jaeger, 1976) have employed a number of different ways of characterizing methods for setting standards. Nearly all reviews identify two basic types of methods -- those requiring judgments about performance levels (item-based methods) and those requiring judgments about performers (examinee-based methods). We also discuss a third type of method -- those where judgments are made about outcomes of pass/fail decisions (outcome-based methods). In the following sections, we review specific techniques that comprise the three, more general categories of standard setting judgment paradigms.

Item-Based Methods

Nedelsky's method

Nedelsky's (1954) method of standard setting can be used only with multiple choice tests. Initially, judges are asked to consider the "minimally competent" examinee. For each multiple choice item, judges are then asked to identify which distractors they feel a minimally competent examinee should be able to eliminate as incorrect. The minimum passing level (MPL) for each item is then defined as the reciprocal of the number of remaining response options, after omitting the options that a minimally competent examinee should be able to identify as incorrect. A standard for each judge is obtained by summing the MPLs across all test items for that judge. A standard for the test is obtained by averaging the individual judge's standards. Nedelsky recommended that test standards be adjusted for measurement error to prevent an acceptable examinee from being failed due solely to measurement error.

Although the Nedelsky method is used frequently to set standards, it has several potential disadvantages. One disadvantage is that use of the method is limited to multiple choice test items. A second more serious disadvantage is that the method makes two unrealistic assumptions (Jaeger & McNulty, 1986). First, it assumes that examinees will randomly choose among options that they cannot eliminate as incorrect. Second, examinees are assumed to have no partial information or to be uninfluenced by partial information when choosing between remaining alternatives. In addition, some

studies (Poggio, Glassnap & Eros, 1981; Meskauskas, 1976) have shown that standards between judges can vary considerably when the Nedelsky Method is employed.

Poggio (1984) has outlined additional potential problems with the Nedelsky method. First, judges have found the method confusing and have reported low confidence in their ratings. Second, the method requires that all judges be highly knowledgeable about the test item difficulties, the job assignments, and the proficiencies of the examinee population. Thus, only judges with particular types of expertise can be used to set standards with this method. Third, judges can be careless in their attention to items, sometimes marking the correct answer as a distractor to be eliminated. Finally, on a practical level, the Nedelsky method typically results in a standard far below that of all other methods. Although Meskauskas (1976) concluded that the method can be used if a sufficient number of judges are able to reach a common consensus, Brennan and Lockwood (1980) questioned the use of Nedelsky's procedure in any situation.

In reference to setting Army job performance standards, the Nedelsky method could only be applied when setting standards on multiple choice knowledge tests. Thus, standards for many of the other performance measures, such as ratings or hands-on tests, could not be set using this method (Jaeger & McNulty, 1986). Another potential problem is that of identifying an appropriate referent population for a "minimally competent examinee." (This, of course, is an issue that is relevant to any technique requiring consideration of minimum competence.) Jaeger and McNulty (1986) have suggested that in a military setting, the referent population could be task specific (i.e. concerned with borderline performance on a specific task or set of tasks) or could refer to an entire MOS (i.e. concerned with borderline performance in the MOS in general). Unfortunately, no research has been conducted to investigate which of these referent groups or possibly others produce the smallest variation in recommended standards. As a final point, it may be impractical to attempt setting multiple performance standards using the Nedelsky method. If judges have difficulty reaching consensus for one standard, large variability around multiple standards may yield distributions of recommended standards that are highly overlapping. It may thus be difficult to clearly distinguish between standards reflecting different performance levels.

Angoff's method

In using Angoff's (1971) method, judges are asked to estimate the probability that a "minimally competent person" would answer each of several dichotomously scored items correctly. In essence, judges must consider a group of minimally competent examinees (not just one) and estimate what proportion of these individuals would answer each item correctly. These proportions are treated as estimates of the probability that an individual minimally-competent examinee will pass the item. A passing standard is obtained for each judge by summing his or her probability estimates across all items. The individual judge standards are then averaged to obtain the overall passing standard (Livingston & Zieky, 1982).

Because the Angoff Method is easy to explain and implement, it is preferred in many situations (Norcini, Lipner & Langdon, 1987; Poggio, 1984; Shepard, 1980). Further, data obtained from this method have been shown to have reasonable psychometric properties (Norcini et. al, 1987). However, a potential disadvantage of the Angoff method, like all methods that require consideration of minimal competence, is that judges may have difficulty agreeing on the definition of a "minimally competent" examinee. Accordingly, each judge may establish his or her own level of scoring, possibly creating variability among different judges' standards (Poggio, 1984). A final limitation of the Angoff method is that its use is restricted to dichotomously scored items.

Despite potential problems, the Angoff method is one of the most commonly used standard setting techniques. This method could likely be applied to setting multiple performance standards in the Army for dichotomously scored test items (e.g., job knowledge test items and hands-on steps that are scored go/no-go). Some adaptation would be required for ratings and other continuous performance measures.

Ebel's method

The Ebel (1972) procedure also asks judges to begin by conceptualizing a minimally competent examinee. Each judge then develops a two-dimensional matrix, with the dimensions labeled "difficulty" of the items and "relevance" of the items. The actual number of item difficulty and item relevance categories is up to the standard

setter, but Ebel suggests three levels of difficulty ("easy," "medium," and "hard") and four levels of relevance ("essential," "important," "acceptable," and "questionable").

Once the matrix is drawn, judges independently place each of several dichotomously scored test items into a cell, based on their expert judgement about that item's difficulty and relevance. Livingston and Zieky (1982) then recommend having the judges discuss their placements, with the opportunity for changing them based on this discussion. However, agreement at this stage is not a requirement. Next, for each category in the matrix (e.g., easy and essential), each judge answers the following question:

> "If a borderline test-taker had to answer a large number of questions like these, what percentage would he or she answer correctly?

Once the percentages have been placed in each cell (and agreed upon by the judges), they are multiplied by the number of items each judge allocated to that cell. The sum of the products across the cells creates the recommended standard for each judge. The recommendations of all judges are then averaged to produce the final recommended standard.

The main advantages of the Ebel method are its ease of implementation and the ease with which judges can understand their tasks. However, there are several potential disadvantages associated with the method (Poggio, 1984). First, the technique is time consuming and consequently, fatigue and boredom can become a problem. Second, judges have experienced great difficulty estimating the percent of minimally competent persons who would pass items placed in the "questionable" category. This problem could conceivably be eliminated by simply disregarding items in this category. A third potential disadvantage is that standards set with this method can vary considerably, depending upon whether they are based on independent judge's ratings or on group values. Like the Angoff method, the Ebel method is also restricted for use with dichotomously scored items.

In addition to the aforementioned limitations, other authors have identified potential problems concerning more global aspects of the Ebel method. First, Hambleton and Eignor (1980) have criticized the fact that Ebel does not prescribe a

precise number of categories to be used in the matrix. These authors maintain that the use of different categories or different numbers of categories is likely to produce variable standards even when applied to identical items. Another criticism was leveled by Meskauskas (1976), who suggested that since judges are more knowledgeable about the material than the minimally competent person, they are likely to ignore fine discriminations that examinees need to make among items in order to answer them correctly. As a result, Ebel's method may result in a higher standard than other methods.

Applying the Ebel method to Army performance standards may be difficult. In addition to the fact that the method is restricted to dichotomously scored items, the number of judgments that would be required to set multiple standards may make use of the method prohibitively time consuming. Another concern is that the Ebel method may yield higher standards than would be obtained from other methods. Perhaps the most serious potential problem, however, is that the Ebel method presumes that all test items are unidimensional but can be stratified on difficulty and relevance dimensions. Jaeger and McNulty (1986) have noted that many military job performance measures do not lend themselves to the type of stratification required in using the Ebel method. If it is not possible to identify relatively homogeneous stratification dimensions or clusters, Ebel's method will likely yield very unstable results.

Jaeger's method

The Jaeger (1982) method was developed for a high school competency test but can be adapted to any situation where a decision is based on test performance. Here, one or more populations of judges must be identified, and then representative samples of judges must be drawn from these. Unlike other procedures where judges are asked to conceptualize a "minimally competent" examinee, the Jaeger procedure asks judges the following question:

> "Should <u>every</u> examinee in the population of those who receive favorable action on the decision that underlies use of the test be able to answer the test item correctly?" (Jaeger and McNulty, 1986)

An initial standard is computed for each judge by summing the number of items to which the judge responds "yes" to the above question. The median of the individual

judge standards is then computed to arrive at an initial test standard. Once initial standards are developed, judges are given several opportunities to reconsider their decisions, based on actual test data and discussion with fellow judges. Jaeger (1982) recommends that the final test standard be the lowest of the recommended standards from all groups of judges.

One potential advantage of Jaeger's (1982) method is that it does not require a judge to conceptualize a "minimally competent examinee", thereby eliminating the need to identify an appropriate referent population and alleviating the possibility that different judges may employ variable standards in defining minimum competence. However, Jaeger's (1982) method does not eliminate the possibility that different standards will influence the judgment process. It is quite likely, for example, that judges will have differences of opinion concerning which items individuals who receive favorable personnel actions should be able to answer. Jaeger (1982) attempts to handle this problem by having judges discuss their standards. While this is a viable strategy, care must be taken to ensure that standards resulting from group discussion do not merely reflect the opinions of the most vocal or persuasive judges (Brennan & Lockwood, 1980). A final potential disadvantage of the Jaeger method is that in certain public schools settings, it was shown to yield unacceptably high standards (Jaeger & McNulty, 1986).

Regarding the application of Jaeger's method to setting Army performance standards, one distinct advantage is that its use is not restricted to written tests. While the Angoff and Ebel methods can conceptually be adapted to non-written situations, they are basically intended for, and applied to, paper-and-pencil tests containing dichotomously scored items. Nedelsky's method is even harder to adapt to a non-written exam, as it is intended only for multiple choice tests. By contrast, Jaeger's method can be applied easily to hands-on or work sample tests, (Shikiar & Saari, 1985). However, some adaptation would be required for ratings or other continuously scored measures. Although the Jaeger method appears to fulfill some necessary criteria for application to Army standard setting, the possibility that it may yield impractically high standards could preclude its use. However, since this method has not been applied to or evaluated in conjunction with performance measures, the extent to which it may be a useful method for setting Army standards is yet undetermined.

Examinee-Based Methods

Borderline-group method

Zieky and Livingston's (1977) borderline-group procedure and contrasting groups method both require judgments about specific examinees rather than test items. The borderline-group method rests on the assumption that the cut score should be based on actual examinees who are borderline, i.e., minimally competent (Livingston & Zieky, 1982). Once selected, judges are tasked with identifying examinees who are competent, borderline, and incompetent. The cut score is then set at the median score of the borderline group. The median score is used because it is less affected by outlying scores. The cut score is commonly reduced slightly to reflect measurement error. In applying this method, there should be relatively small variance in the scores of the borderline-group members. Widely varying scores suggest that the judges may have: a) identified many test takers as borderline who do not belong in that category; b) based their judgments on something other than what the test is measuring; or c) differed substantially in their own standards for the examinees (Livingston & Zieky, 1982).

Contrasting-group method

This method assumes that examinees can be divided into a qualified group and an unqualified group (Livingston & Zieky, 1982). Judges are asked to identify students they are certain have mastered the material and those who they are certain have not mastered the material. The score distributions are then plotted and the initial cut score is the point of intersection between the two distributions. Like the Borderline-Group method, the final cut score can then be adjusted to reflect measurement error.

The main advantage of the Borderline-Group and Contrasting-Group methods is their relative simplicity. These techniques are easy to explain and implement (Livingston & Zieky, 1982; Poggio, 1984). In addition, categorizing actual people as masters or nonmasters may be a relatively easy judgment for experts who rate their employees or students often and are familiar with their performance (Jaeger & McNulty, 1986). Another advantage of examinee based methods is that they are not restricted for use with particular types of items or lists. Rather, these procedures can

be used with all types of measures, regardless of how they are scored. One potential disadvantage of these methods is that they require defining groups of definite masters and non-masters (Jaeger & McNulty, 1986), which can be difficult judgments to make. While judges have not reported serious problems with the contrasting groups method, it can be difficult for judges using the borderline-group method to identify the typically small percentage of test takers who are truly borderline (Livingston & Zieky, 1982; Poggio, 1984).

Jaeger and McNulty (1986) have discussed other threats to the validity of examinee-based methods. First, if the sample of examinees used is not representative of the population of examinees to which the standard is to be applied, a biased standard will result. Second, examinee-based methods require that judges be familiar with examinees' performance. However, familiarity with examinees increases the likelihood that judgments will be influenced by halo, a tendency to consider factors other than relevant knowledge or skills in making competency evaluations. A final potential problem relevant to the borderline group method is that judges who are less confident of their ratings may commit central-tendency error. That is, uncertainty may lead them to place examinees in the "borderline" category simply to avoid the extreme "competent" and "incompetent" ratings.

The Contrasting Group and Borderline Group methods have some advantages and some disadvantages for setting performance standards in the Army. As mentioned, these methods can be used to set standards on tests that are neither item-based nor scored dichotomously. In addition, examinee-based judgments are consistent with subordinate performance judgments routinely made by Army supervisors. One disadvantage of applying examinee based methods to the Army is that they cannot easily be used for setting performance standards for new MOS in which job incumbents do not yet exist. A second potential disadvantage (similar to the item based methods discussed previously) is that an appropriate referent population (e.g., task-based or MOS-based) for setting the performance standards would need to be identified. The third and perhaps most serious problem with these methods is that it may be extremely difficult to find examinees who are considered "unacceptable" (Jaeger & McNulty, 1986). This presents a potential problem because obtaining a stable distribution of test scores is contingent upon identifying a sufficiently large sample of "unacceptable" examinees.

Due to the fact that soldiers within an MOS have been extensively screened, very few unacceptable performers may exist within these jobs.

Outcome-Based Methods

In the preceding two sections, the most commonly used methods for setting standards were reviewed. There is a body of literature on decision theory, however, that has generally not been included in most discussions of standard setting. Since standards generally reflect operational decisions that may be made (e.g. terminate, retain, or promote), the decision theory literature is potentially relevant. A decision theory approach would go beyond either item-based or examinee-based methods and would focus on an evaluation of outcomes associated with performance at different levels. We include here a brief summary of such an approach as it may relate to standard setting.

Most of the basic concepts of decision theory have been outlined by Edwards (1971), Gardiner and Edwards (1975), Keeney (1972) and Raiffa (1968). The decision theory approach involves enumeration of outcomes or consequences of each decision alternative. For example, in evaluating a clerk typist whose performance is marginal, the decision to retain the employee could lead to delays in producing documents and errors in the documents that are produced. Resources might be needed for additional supervision and training. The decision to terminate the employee, on the other hand, would involve consequences of staff shortages and/or replacement and retraining costs.

The basic step involved in employing a decision theory approach, as outlined by Edwards, are:

1. Identify the individuals or units who have a stake in the outcome of the decision.

2. Identify the decision(s) to be made.

3. Identify the outcomes (options) to be evaluated.

4. Identify the relevant dimensions of value (consequences).

5. Estimate importance weights for the different dimensions.

6. Measure the location of each outcome on each of the dimensions of value.

7. Calculate the weighted utilities (sum the products of the value dimension amounts and importance weights for each outcome).

8. Select the largest.

Decision theory was originally designed for decisions where there were a number of similar alternatives, such as buying a car or selecting a new employee among several applicants. The extension of decision theory to binary decisions such as whether or not to retain, whether or not to retrain, and whether or not to promote an employee, requires some extension of the general decision theory paradigm. The potential advantage of such an approach is that it would provide a more detailed rationale for particular standards. Specific consequences of good and bad performance would be enumerated and evaluated. The final standard would be linked to these specific consequences.

While discussion of these consequences would still necessarily involve judgments, there is at least some possibility for empirical support. In the clerk-typist example, empirical data might be used to estimate the extent of delays associated with various typing speeds and error rates. In some instances, it may be possible to assess dollar consequences of specific outcomes.

Our initial review of the literature did not reveal much on the use of decision theory in setting performance standards. Decision theory models have been used in linking selection standards to performance outcomes as described below (pages 39-41).

Other Methods

In this section, we will briefly describe several other methods that are available to the standard setter. Two of these methods compare standards of performance by

evaluating subsequent classification errors. This is in contrast to the previous methods which focused on test content (Meskauskas, 1976). First, Berk (1976) has suggested a method similar to that of the contrasting-groups method. He uses empirical data of "instructed" and "uninstructed" students and suggests three ways to set standards based on these data: 1) classification of outcome probabilities; 2) computation of a validity coefficient, and 3) a utility analysis (Hambleton & Eigner, 1979). While the goal of minimizing classification errors is not unusual in standard setting methods, Berk's method is more easily understood and implemented than other such approaches. A major concern with this method is that it essentially equates "instructed" with "competent". All Army job incumbents have been through training. The goal in setting standards is to identify "instructed" soldiers who are still not performing competently.

Another method is the Kriewall (1972) model in which students are classified as non-master, master, or in-between (analogous to a borderline classification). The model, which focuses on identifying the likelihood of classification errors, is based on a binomial distribution which requires several assumptions, such as a randomly selected group of dichotomously scored items and independent responses to questions. Boundary values are set, an initial cut score is decided upon, and the probabilities of misclassification errors are estimated based upon the cut score. Actual data are not required for this model, which can be advantageous (Hambleton & Eignor, 1979). However, Kriewall's method requires satisfying several assumptions and is also very complicated to employ. Thus, the method may not be suitable for setting Army performance standards.

A fairly new method has been proposed by Cangelosi (1984), who suggested establishing a cut score concurrently with the development of test objectives. The persons developing the objectives also specify the proportion of correct answers a borderline student would be expected to achieve for items representing each objective. The cut score is then the weighted sum of the expected proportions for all objectives on the test. Cangelosi (1984) argues that the method's main advantage is that standard setters must define "success" early in the test development process; hence, the test may be more valid. Research by Saunders and Helsley (1987) suggests that Cangelosi underestimated the difficulties associated with his method, which primarily resulted from failing to consider individual test-item difficulties. The authors provide data indicating that the Cangelosi (1984) method yields highly inconsistent results and

therefore may be of little use. In addition, other authors (Livingston, 1982; MacPherson, 1981) have found that using test developers to set standards yields higher standards than when other groups of judges are employed. Although Cangelosi's judges were developing test objectives rather than test items, it is possible that this involvement may have affected the recommended standard in ways that would not be consistent with judges who were not involved in developing the objectives.

Comparison of Methods

Several studies have been undertaken to empirically compare various standard setting methods. The overwhelming finding to emerge from the body of research is that different or even the same standard setting methods frequently yield widely discrepant results (e.g., Andrew & Hecht, 1976, Brennan & Lockwood, 1980; Halpin & Halpin, 1983; Koffler, 1980; Livingston & Kastrinos, 1982; Livingston & Zieky, 1983; Sigmon & Halpin, 1984; Skakun & Kling, 1980). For example, Glass (1978) noted that cut scores found in Andrew and Hecht's (1976) comparison of the Ebel and Nedelsky models would have corresponded to pass rates of 50% and 95%, respectively. Similarly, in comparing the Nedelsky and Contrasting Group approaches, Koffler (1980) found that the two methods produced discrepant standards on three of eight tests. Further, there was no consistent pattern of agreement between cut scores generated by the two methods. As a final example, in evaluating the reliability of the Nedelsky method, Livingston and Kastrinos (1982) found large variations among judges as well as a higher standard when judges performed the task a second time.

Results of research comparing various standard setting methods raise questions about the confidence that can be placed in a given cut score. On the other hand, the extent to which variable standards should be of concern has been a debated issue among researchers. Several authors have argued that because standard setting is, by definition, judgmental, variable standards should not be unexpected nor should they be cause for concern (Andrew & Hecht, 1976; Hambleton, 1978). Glass (1978) however, takes issue with this viewpoint. In particular, he argues that when different standard setting procedures have identical purposes and are designed around the same conceptualization of minimum competency (e.g., like in Andrew & Hecht, 1976), then they should result in similar standards. The fact that significantly different standards are often observed suggests that the technique used to set standards is the most important

determinant of that standard, raising serious questions about the validity and utility of the techniques employed. Yet other authors (Scriven, 1978; Block 1978) have maintained that irrespective of potential problems associated with the standard setting techniques, flawed standards are better than no standards at all.

Given that there are obvious differences in the standards yielded by various standard setting methods, questions concerning which methods are most effective become increasingly important. Berk (1976) suggested six criteria for evaluating the effectiveness of different standard setting procedures. These are:

1) the method should correctly classify examinees;
2) the method should be sensitive to different levels of examinee performance;
3) the method should be sensitive to instruction or training;
4) the method should be psychometrically and statistically sound;
5) the method should identify the true standard; and,
6) the method should produce validity evidence, given the importance of defending decisions made based on the standard.

For the methods discussed here, the second, third, and sixth criteria are difficult to meet without actual performance data.

Berk (1986) reviewed many empirical studies that compared the standard setting methods in various combinations. He concluded that the Angoff method offers the best mix of technical adequacy and applicability, even given the difficulties of its meeting some criteria. Hambleton and Eignor (1979) and Shepard (1980) also favored the Angoff method because of its simplicity. The Contrasting Groups method was also rated highly by Berk, due to its technical adequacy.

Summary

In this section, three general categories of standard setting judgment paradigms have been reviewed: item-based methods, examinee-based methods, and outcome-based methods. Although all three general method types have, at least, some potential applicability for setting performance standards, an outcome-based approach has the advantage of providing more explicit rationales for whatever standards are set.

Some item-based methods could be used to set performance standards, with measures composed of dichotomously scored items. Of the methods reviewed here, Angoff's appears to be a primary candidate for setting performance standards, due to its technical adequacy yet relative simplicity (Berk, 1986). The existing item-based methods do not, however, hold much promise for use in conjunction with continuously scored performance measures (e.g., rating scales). For these types of measures, it would be necessary to modify existing item-based methods extensively or develop entirely new standard setting methods. As an example, it might be possible to set standards for behaviorally-based rating scales (such as those used in Project A mentioned on page 4) by having judges sort a sample of critical incidents (on which the scales were based) into two categories titled "below minimum competency behaviors" and "above minimum competency behaviors." During the scale development process, a sample of subject matter experts (SMEs) rated the effectiveness of each critical incident. A mean effectiveness rating was then computed for each incident by averaging across individual SME ratings. By comparing the distributions of mean effectiveness ratings of incidents placed in the "below minimum competency" versus "above minimum competency" categories, it should be possible to identify a fairly precise scale value that could serve as the minimum competency standard. For instance, the effectiveness level of the point of intersection between the two distributions of incidents may serve as an initial cut score. As an additional point, it would likely be possible to develop multiple performance standards using procedures like these.

Regarding examinee-based methods, the most potentially serious problem here is a lack of "unacceptable" performers which are necessary to obtain a stable distribution of scores on the measure(s) of interest. Under most circumstances, poor performers will either self-select out of their jobs or be terminated by the organization (Schneider & Schmitt, 1987). On the other hand, exceptional performers may be promoted quickly. Given that extensive screening takes place prior to placing recruits into an MOS and then again prior to graduating them from military service schools, identifying poor performers may be a particularly serious problem in the Army. Thus, unlike many testing situations in which an adequate distribution of scores can be obtained, distributions of performance scores are more likely to suffer from range restriction. Finally, even if restriction of range were not a problem in the Army, examinee-based methods could not be used to set standards in new MOS that do not yet have job incumbents.

In such cases, standards would have to be adapted from those set for similar jobs, requiring a number of assumptions about the generalizability of standards.

THE JUDGMENT PROCESS

The judgment process has been frequently identified as the foundation of standard setting. In addition to standard setting methods, variables that can influence the judgment process and resulting standard include:

- Judgment facilitation techniques (e.g. use of normative data, Delphi techniques, or other iterative processes);
- Judge characteristics;
- Judge training; and,
- Number of judges.

Judgment Facilitation Techniques

In an attempt to improve the judgment process involved in setting standards, judgment facilitation techniques have been introduced, including the use of normative data and iterative judgment processes. These judgment facilitation techniques can be used with virtually any standard setting method and are reviewed in the following sections.

Normative data

The use of normative data has been recommended in standard setting situations where decisions are based on knowledge of examinee capabilities as well as job requirements (Jaeger & McNulty, 1986). An example of such a standard setting situation would be one in which on-the-job training may compensate for marginal qualifications at hiring time. Normative data concerning examinees' test performance allow judges to evaluate the consequences of their recommended test standards. In addition, normative data provide judges with information which seems to aid them in making more educated recommendations for appropriate test standard levels. Finally, the use of normative data has been shown to reduce the variability of judges' standards, and in turn, increase the reliability of judgments (Cross, Inpara, Frary, & Jaeger, 1984; Jaeger & Busch, 1984).

Hambleton and Powell (1983) believe the decision whether or not to use normative data should be dependent on the goals and constraints of the testing program. Specifically, these authors argue that although the use of normative data makes the standard setting task easier, most standard setting situations are not concerned with the status quo of the examinee population. It could thus be a mistake to use normative data exclusively in establishing performance standards. This is because by doing so, emphasis is shifted from setting a standard based upon "what should be" to one based on "what is." Nevertheless, many researchers have recommended the use of normative data as a reality check on judges' recommended standards (Hambleton, 1980; Jaeger, 1978; Shepard, 1980).

The use of normative data may prove to be an important factor for setting performance standards in the Army. An optimal use of such data would likely be as a reality check mechanism. Judges could consult normative data either before or after making their recommendations and use this information to ensure that realistic standards were set. It is also likely that the use of normative data would help to increase consensus among the standard setting judges.

Iterative judgment processes

Iterative techniques have also been used to facilitate the judgment process (Jaeger, 1982). The general approach is to first compute an initial standard for each judge. Then, all judges are given opportunities to reconsider their initial recommendation, using the recommendations made by all other judges. Iterative judgment processes can be used alone or in conjunction with normative data. Regarding the feasibility of implementing interactive techniques, Jaeger (1982) found that they posed minimal practical difficulty and could be completed within a reasonable amount of time. However, Berk (1986) has also warned that iterative judgment processes can be quite tedious and expensive to employ.

The Delphi technique, which is a variate of the iterative approach outlined above, was originated by Dalkey (1969) as a method to measure group opinion. Judgments are first made independently and anonymously. The judgments are then pooled, summarized, and fed back to the judges for another round of opinion. At this point, judges are typically allowed to discuss their recommendations and present rationales for

their ratings. Jaeger and Busch (1984a) investigated the effects of a Delphi modification of the Angoff technique to set standards for National Teacher Examinations. In addition to employing a Delphi Procedure, the judges were also provided with normative data. Results of this investigation showed that the use of normative data combined with the Delphi process led to reduced variability among recommended standards but did not seem to have a significant effect upon the mean recommended standard.

Although iterative techniques can be advantageous, a cautionary note regarding their use is warranted. If data concerning judges' ratings are provided without justification, this can lead to a shift in judgment toward central tendency of the group. On the other hand, if judges are allowed to provide rationales for their ratings, the most vocal individuals are likely to control the discussion, possibly inappropriately influencing the remainder of the group. It is thus recommended that when judges provide their ratings, they justify these recommendations in a controlled discussion format which should, in turn, result in better informed, less biased judgments.

When setting Army job performance standards, some type of iterative judgment processes could be employed. Empirical results have indicated that a controlled iterative discussion process can lead to less variability among judges and better informed standard setting decisions. However, if such an iterative process is implemented, it will be crucial that discussions are carefully monitored so that no one judge or group of judges can dominate the process and consequently, bias decision making for the entire group.

Judge Characteristics

Meskauskas (1983) has suggested that the choice of judges may impact the standard setting process as much as the choice of procedures. In this section, we will first review general guidelines and suggestions made by various authors for selecting judges to participate in setting standards. Following this discussion, empirical research investigating judge characteristics will be reviewed, although it should be noted that the number of empirical studies conducted in this area has been relatively small.

In their review of the literature, Hambleton and Powell (1983) provided a set of questions to consider when choosing judges for standard setting. The questions include:

1. Which demographic variables should be used in selecting judges?
2. How should names of possible judges be generated?
3. Which individuals should be involved in the judge selection process (and why)?
4. How many judges should be selected to participate?
5. Should judges volunteer or should they be conscripted?
6. Should judges be selected to be representative of some constituency?
7. Should "expert" judges be preferred over representatives of groups of interest?
8. When judges are arranged into working groups, what is the optimal group size, and should the groups be formed homogeneously or heterogeneously?
9. Should data from judges be discarded when there is reason to believe that they were unqualified to do the job, or carried out the task in a "sloppy" fashion? Should specific steps be taken to identify "poor" judges?
10. Should judges be paid for their time?

For several of the questions listed above, Hambleton and Powell (1983) proceeded to provide more specific recommendations. Regarding question 1, judge demographics, the authors maintain that variables such as race, sex, age, level of education, occupation, specialty, and willingness to participate can serve as potential influencing variables in the judgment process. The authors also point out that the composition of the standard setting committee is often crucial for lending credibility to the resultant standard. Thus, depending on the particular characteristics of the situation, the importance of different demographic variables is likely to vary.

When deciding who should be involved in selecting judges (question 3 above), Hambleton and Powell (1983) recommend that individuals representative of all constituencies or "interest" groups be included in the judge selection process. This is done to prevent alienating any group(s) from the total standard setting process, thereby increasing the chances that the resultant standard(s) will be accepted.

In reference to question 7, the authors suggest that whenever possible, panels of judges should be composed of both "experts" and representatives of different stakeholder groups. Finally, question 8 concerns issues relevant to dividing judges into working groups. Hambleton and Powell (1983) outline three conditions under which working groups should be formed. These are:

- When there is interest in promoting discussion but the entire group of judges is too large to permit effective discussion.
- When there is interest in comparing the standards generated across similar groups (i.e., assessing the reliability of standards).
- When there is interest in comparing the standards generated across dissimilar groups (i.e., assessing the validity of standards).

Thus, the goals of the particular standard setting process should dictate whether or not working groups are formed as well as the composition (homogeneous vs. heterogenous) of those groups.

Jaeger and McNulty (1986) provided additional recommendations concerning the choice of judges. When using item-based standard setting methods, judges should be knowledgeable of examinee skills and abilities measured by the test, as well as of the distribution of examinees' performance across each test item. For examinee-based methods, which require classification decisions, judges must be knowledgeable about each individual's competence or degree of mastery in the particular subject area(s).

Although empirical research investigating judge characteristics has been relatively limited, some interesting findings have been reported. For example, Livingston (1982) found that for the Ebel, Angoff, and Nedelsky methods, it is a mistake to use the individuals who constructed the test as judges because they set a higher passing score than judges who were not involved in test development. Similarly, MacPherson (1981) examined test scores for subordinates of Non-Commissioned Officers (NCOs) who were involved in test construction and also in training. He found that the subordinates of these NCOs had relatively low examination scores and attributed this to the NCOs

underestimating the item difficulties. Both of these investigations thus suggest that involvement in the test development process may lead judges to set higher standards.

Other research conducted by Jaeger (1982) examined the degree of similarity of standards recommended by judges having different interests and involvement in secondary education. In addition, this investigation examined the relationship between certain background characteristics of judges and the standards they recommended. Results from this investigation indicated that different types of judges will likely disagree on recommended standards. For example, registered voters felt that current test standards for North Carolina High School competency tests were too lenient, while high school teachers wanted the standard on the reading test to be lowered and school principals and counselors wanted it raised somewhat. There was majority agreement across all types of judges that the mathematics test standard should be raised; however, there was not agreement concerning how much it should be raised. Findings concerning the relationship between judge demographic characteristics and standard setting (e.g. sex, race, age, parent, years of education, children in high school, children who took competency test), revealed that the only demographic variable that was significantly predictive of judges' recommended standards was race. The mean recommended standard of black judges was approximately twelve points lower than that of white judges.

In summary, several factors (e.g., demographics, group membership) have been purported as important considerations in selecting standard setting judges. However, we could find only two empirical investigations (McPherson, 1981; Jaeger, 1982) that investigate the relationships between judge characteristics and standards. These investigations suggest that group membership, race, and whether or not the judge was involved in test development are important determinates of recommended standards. Future research should be conducted to investigate whether or not the above findings generalize to other standard setting situations, the conditions under which various demographic variables may be important, and <u>why</u> different referent groups disagree on recommended standards (i.e., do different referent groups use different information or weight the same information differently in determining a standard?).

Since no research investigating judge characteristics has been conducted in a military setting, it is difficult to specify what the optimum mix of judges should be for

the Army and the precise characteristics these individuals should possess. In addition, selection of appropriate judges may depend, at least in part, on the standard setting method used. Use of item-based methods, for example, would require judges to be knowledgeable about the distribution of examinees on the measures of interest. Use of examinee-based methods would require the judges to be knowledgeable about the actual job performance of the soldiers they are classifying. Regarding possible "interest" groups that should be represented in the standard setting process, it would seem reasonable to include a representative mix of judges balancing detailed job knowledge against overall responsibility for Army policy.

Judge Training

Several authors, (e.g., Jaeger & McNulty, 1986; Jaeger & Busch, 1984a) have discussed the importance of training judges to perform their standard setting task. Although training is likely to vary as a function of the particular standard setting paradigm used, Jaeger and McNulty (1986) have identified some common themes. First, the authors discuss the importance of familiarizing judges with the test for which they will be setting standards. One technique for accomplishing this has been to have the judges actually complete the test of interest under conditions that approximate an operational testing environment (Cross, et al., 1984; Jaeger, 1982).

Second, Jaeger and McNulty (1986) maintain that judges must understand the sequence of operations they are required to perform in recommending standards. For some standard setting procedures (e.g., those requiring a single set of recommendations), the operations required are uncomplicated and relatively easy to teach. Other standard setting procedures (e.g., iterative judgment processes requiring multiple sets of recommendations) are significantly more complex and thus require more extensive training. As one example, Jaeger and Busch (1984a) used a simulation of the judgment process, along with a simulated version of the test for which standards would be generated, to train their judges to perform a three-stage standard setting operation. Although the effects of training were not directly evaluated, judges reported clearly understanding what they were to do subsequent to participating in the simulation.

Finally, Jaeger and McNulty (1986) argue that when judges are provided with normative data reflecting examinees' test performance, they must be trained how to

properly interpret these data. If, for example, judges are provided with estimated difficulty values ("p-values") for each item on a test, it should not be assumed that judges will know what these numbers mean. Thus, the meaning of a "p-value" should be explained. Similarly, if graphs or frequency distributions are presented, judges should be taught how to properly read and interpret these materials.

Although it might be expected that the above described types of training may well serve to increase the degree of consensus among judges' recommended standards, we are unaware of even one investigation that has empirically evaluated training effectiveness in a standard setting situation. Another related issue concerns whether or not certain types of training are more effective than others in increasing judge agreement, and if so, why? Again, we are aware of no empirical research addressing this issue for a standard setting application. However, some rater training research has been conducted in the performance appraisal and employment interviewing areas that may have relevance for the standard setting process. This literature is reviewed below.

Similar to arguments made by Jaeger and McNulty (1986), performance appraisal researchers (e.g., DiNisi, Cafferty, & Meglino, 1984; Feldman, 1986) have recognized the importance of familiarizing raters with their rating task as well as the procedures and operations required to perform it. In fact, empirical research has shown that when trainees are provided with proper training (i.e., training that is based on and congruent with the rating task demands), rating quality increases. For example, Pulakos (1984, 1986) developed training that focused on teaching raters what specific types of data should be attended to, how these data should be interpreted, and importantly, how these data should be used in formulating the particular judgment required by a given rating task. This training was shown to yield significantly more reliable (higher inter-rater agreement) and accurate (valid) ratings than no training or "incongruent" training (i.e., training not developed in accordance with the particular rating task demands).

Another type of training, rater "error" training, might be applicable to some standard setting situations, especially ones in which examinee-based methods are employed. This training was initially developed by Latham, Wexley, and Purcell (1975) to train employment interviewers to reduce several common rating errors (e.g., halo, central tendency) in their evaluations. Many authors have since developed and evaluated a variety of error training programs (e.g., Bernardin, 1978; Bernardin &

Walter, 1977; Borman, 1975), both in interviewing and performance appraisal contexts. These programs have generally been quite effective in reducing common rating errors.

In summary, the research reviewed here suggests that some form of training probably should be provided to standard setting judges. At a minimum, this training might be focused on familiarizing judges with the measure(s) for which standards are to be set, ensuring that judges understand the sequence of operations they will be expected to perform, and providing explanations regarding how to interpret any normative data to be used. In addition, Pulakos (1984, 1986) indicated that it may be possible to develop training that focuses directly on improving standard setters' ability to make the particular judgments required by a given standard setting procedure. Or, by using procedures similar to those described by Latham et al. (1975), it may be possible to eliminate the halo and central tendency effects that are often associated with use of examinee-based methods. Finally, it should be noted that irrespective of the type of training administered, it is likely to be more effective to the extent that trainees are provided with opportunities to practice and receive feedback on their judgments (Goldstein, 1986; Wexley & Latham, 1981).

Number of Judges

Not only is it important to choose judges carefully and provide them with proper training, but consideration must also be given to determining the optimal number of judges to select. When too few judges are used in setting standards, there is a risk that the resulting standard error of the recommended test standard will be large. Of course, use of too many judges may waste resources and unnecessarily prolong or complicate the standard setting process. The ideal number of judges is that which reduces the standard error of the recommended test standard to less than half a raw score point on the test (Jaeger & McNulty, 1986). However, it is not always practically feasible (e.g., because of cost or judge availability) to obtain sufficient sample sizes that would reduce the standard error of the recommended test standard to the ideal level. An alternative approach uses the relative magnitudes of the standard error of the recommended test standard and the standard error of measurement of the test for which a standard is desired as a basis for determining how many judges should be used (see Jaeger and McNulty, 1986 for a detailed explanation of this method).

The standard error of our performance standards will, of course, depend on the degree of consensus achieved by a particular method, as well as on the number of judges. Also, when judges represent various interest groups, the number of different groups will be important. Unfortunately, relatively few studies (for exceptions, see Cross, et al., 1984; Jaeger and Busch, 1984a) have investigated the optimum number of judges required for different judgment paradigms. Beyond the standard setting method itself, however, it is also likely that judge characteristics (e.g., education level, degrees of "expert" knowledge, number of constituencies represented) and standard setting process characteristics (e.g., type and amount of training, type and amount of normative information provided) may affect the degree of consensus achieved and thus the number of judges required. Thus, research should be conducted to estimate the accuracy that can be attained with particular combinations of these variables, method, and judge sample size. Finally, on a more practical level, it should be noted that accuracy goals may not be the only important determinant of the required number of judges. The need for allowing participation by multiple constituencies may further increase sample sizes.

Summary

Improving on the judgment process has been identified by many measurement specialists as the key to improving the objectivity and accuracy of standard setting procedures. By carefully selecting a knowledgeable and representative sample of judges, the validity, reliability, and potential for acceptance of the resultant standard can be increased. Careful training of the judges and the use of iterative judgment processes can also increase the quality of the standards obtained. Variation in judgments can be reduced, consequently increasing the quality of the standards. Finally, providing normative data to judges can increase the reliability of judgments and also allow judges the opportunity to evaluate the consequences of their recommended standards.

Many researchers have incorporated some or all of the above recommendations for facilitating judgment tasks and consequently, improvements in the standard setting process have been reported. The Army could likewise benefit by incorporating these suggestions into the design of the performance standard setting process. Judges chosen to help set Army performance standards should be carefully selected based upon prin-

ciples of representation and expertise. The appropriate number of judges should be determined, and judges should be adequately trained in the standard setting process. It is feasible that the standard setting methodology used for setting Army performance standards will include such factors as the provision of normative data and an iterative judgment process involving feedback. Once standards are established, however, it will then be necessary to combine the multiple standards set for individual performance dimensions into an overall job performance standard. Issues surrounding the development of an overall performance standard are discussed in the following section.

COMBINING MULTIPLE STANDARDS

Given that job performance is inherently multidimensional, one challenge associated with the development and use of job performance standards is how to best combine standards set for multiple dimensions into an overall job performance standard. Development of an overall standard is, of course, a necessary prerequisite to linking performance standards to selection standards. The central issue to be considered here is that an employee's job performance may be quite satisfactory in some areas but not satisfactory in others. Thus, decisions must be made regarding the extent to which more effective performance in some areas compensates for less effective performance in others. These decisions will dictate how multiple standards should be combined into an overall performance standard.

The question of how to set an overall standard for job performance must necessarily be preceded by the development of a scale for assessing overall job performance. Several different approaches for developing such an overall performance scale, ranging from a simple linear composite to more complex conjoint measurement techniques, were examined as part of the Project A research (Sadacca, Park & White, 1986). A conjoint measurement approach (e.g., Luce & Tukey, 1964; Green & Srinivasan, 1978) asks judges to evaluate trade-offs among increments and decrements along different dimensions. For example, two soldiers, one having a slightly higher level of proficiency and a slightly lower level of motivation than the other, might be compared in terms of their overall contribution to the organization.

In its general form, the conjoint measurement model would not assume that the value of a performance increment is necessarily the same for different parts of different dimensions. It is possible, for example, that small decrements below minimum levels in some areas are balanced only by large increments above minimum levels in other areas. There are two special cases of interest in setting an overall performance standard. In the first case, no amount of increment in other areas can compensate for below standard performance on any other dimension. Using this model, known as a Multiple Hurdles Model, an examinee fails the overall standard if he or she fails any of the individual standards. The other special case of interest is a strictly linear model, where overall performance is measured by a weighted sum of the individual

performance measures. Using this model, known as a Compensatory Model, a decrement in one performance area could be compensated for by increments in other areas.

The Project A results indicated that linear composites provide reasonable approximations to the conjoint scaling results (Sadacca, et.al., 1986). With respect to standard setting, this suggests a Compensatory Model may be appropriate for combining multiple performance standards into an overall job performance standard. Thus, the overall standard would be a weighted sum of the individual standards set for each dimension. The current Army selection policy employs a combination of the Multiple Hurdle Model and the Compensatory Model to screen applicants. Specifically, applicants must pass a moral screen, which eliminates individuals convicted of certain crimes, and a physical screen, which eliminates those who do not meet Army physical standards. Additionally, applicants must pass overall and specific cognitive ability standards. The cognitive standards are based on composite scores from the ASVAB, a multiple aptitude battery. These composites are formed using a Compensatory Model (e.g., each applicant is given one score on the AFQT, which consists of four ASVAB subtests). In this way, the Army screens recruits for different critical requirements using the multiple hurdle model, although the cognitive requirements are assessed using a compensatory model.

LINKING SELECTION STANDARDS TO PERFORMANCE STANDARDS

The final topic of concern in this review is the linkage of selection standards to performance standards. There are two primary issues in this linkage. The first stems from the lack of perfect prediction of job performance levels using available selection test scores. In the absence of perfect correlation, prediction errors are inevitable. Selection standards must be set on the basis of probabilistic information about the likelihood that performance standards will be passed.

The second issue in linking selection standards to performance standards stems from the interaction with training effects. The level (and cost) of training is not necessarily constant. The costs of selecting more able individuals who will require less training must be traded off against the cost of additional training to assure adequate performance for less able selectees. Alley (1987) provides a discussion on the use of occupational learning difficulty as a basis for selection standards. For the present, however, we see little possibility for estimating performance level probabilities under training conditions different from those currently employed. As a consequence, we will limit the focus of this review to the first issue mentioned above and assume that level of training is held relatively constant.

Dichotomous Linkage Models

Most efforts to take account of prediction error in setting cut-off scores are based on a dichotomous model. In such a model, a single accept-reject decision is made and a single dichotomous outcome (acceptable or unacceptable performance) occurs subsequently. Much of this work stems from a criterion-referenced perspective in which the score for which a cut-off is sought is taken as a fallible measure of the underlying trait that defines the outcome. In this sense, the primary issue is the reliability of the fallible measure. In the context of the Army Synthetic Validation Project, a prediction paradigm is more appropriate. In this case the validity of a selection test composite for predicting subsequent performance is at issue and not just

the reliability of the selection composite. In all other respects, however, the issues and results are the same.

The essential ingredients of dichotomous linkage models have been laid out by a number of authors including, for example, de Gruijter and Hambleton (1984) and van der Linden and Mellenbergh (1977). These ingredients include:

- a joint distribution function giving the probability of possible combinations of (fallible) selection test and (true) performance scores,
- a performance score cut-off that separates acceptable from unacceptable performance (or mastery from non-mastery, etc.), and
- a set of "loss" functions that describe the cost of alternative classification decision outcomes.

In most cases, the joint distribution function is assumed to be bivariate normal with prediction score mean and variance estimated from observed values and correlation taken as the estimated validity of the prediction composite. In a criterion-referenced paradigm, the true score mean and variance are not observed and must be inferred using a Bayesian model (Hambleton and Novick, 1973) or fitting specified models, most commonly a binomial model (de Gruijter and Hambleton, 1984). In a prediction paradigm, the criterion distribution can be estimated empirically based on observed criterion score distributions and information about the reliability of the criterion measure.

The primary focus of research on dichotomous linkage models has been on the form of the loss functions that are employed. Huyhn (1976) and Mellenbergh, Koppelaar, and van der Linden (1977) considered loss functions that are zero for correct classifications and separate constants for each type of classification error.

Van der Linden and Mellenbergh (1977) considered separate linear loss/value functions for relating actual performance to the cost or value of selection and rejection decisions. In this model, selection of an individual who performs well above the minimum would be worth more than selection of an individual performing at the

minimum and selection of an individual far below the minimum would "cost" more than selection of an individual whose performance would be barely below the minimum. Similarly, rejection of an individual who would have performed well above the minimum costs more than rejection of an individual who would have performed at the minimum and rejection of an individual who would have performed well below the minimum "saves" more than rejection of an individual who would perform just below the minimum. Van der Linden and Mellenbergh showed that the use of linear loss functions leads to a relatively simple formula for specifying the optimal selection cut-off score in terms of population distribution and loss function parameters.

De Gruijter and Hambleton (1984) point out further issues in the use of a decision theory approach to setting cut-off scores. The difficulties in specifying utility or loss functions and the problems in estimating true score distributions are chief among the issues that they discuss.

More Complex Linkage Models

Current Army selection procedures employ more than a simple dichotomous selection decision. Applicants must pass an overall selection cut-off based on their AFQT scores and also a separate cut-off score based on a prediction composite for the particular job for which they are applying. In addition, a fixed percentage of the training seats for each job are reserved for "quality" applicants. To qualify for these particular seats, an applicant must be a high school graduate and must pass a higher AFQT cut-off.

The essentially trichotomous selection procedures for each Army job reflect an essentially trichotomous view of performance. There is not only an implicit minimum standard for performance during the first tour, but also a higher standard for soldiers who will continue in the Army and become leaders within each occupation. The quality standards are designed to assure an adequate supply of recruits who will meet these higher performance standards.

Unfortunately, our literature search did not find relevant examples of trichotomous or other more complex linkage models. Some attention has been given to nondichotomous models of performance utility. Sadacca et. al. (1986) have developed

relatively continuous functions for describing the utility of different levels of performance in different Army jobs. Nord and White (1987) point out some key issues in estimating the utility of different levels of performance including the difference between absolute and marginal utility. The value of having another leader in a particular job may not be a constant, but may instead depend on the number of leaders already available.

Summary

The basic information for linking selection and performance standards includes (1) the performance standards, (2) estimates of the population distributions for selection and performance measures, and (3) empirical or synthetic estimates of the validity of the selection composite to be employed. With this information, it is possible to estimate the probability of different levels of job performance for applicants at each selection composite score level.

The procedures reviewed involve tradeoffs between alternative selection errors (false positives and false negatives). An evaluation of the "costs" of each type of error must be considered along with their probabilities. Such evaluations become more complex when performance is considered to be more than a simple dichotomy. None of the procedures identified adequately addresses this complexity.

SUMMARY

Procedures for setting job performance standards are best viewed from a systems perspective. The quality and acceptance of standards that are developed depend on an interaction among detailed procedures employed, the characteristics and training of the judges, the types of measures for which standards are to be set, and the overall purpose for setting the standards. At the outset, we proposed a model (illustrated in Figure 1) that describes major interactions among these different components. Research on the development of job performance standards for the Army must consider each of the components in this model.

Three general types of procedures were reviewed: item-based methods, examinee-based methods, and outcome-based methods. Item-based methods appear to offer promise for many types of performance measures, including job knowledge tests and dichotomously scored hands-on performance steps, but may be difficult to apply to more global measures (e.g., ratings). The use of examinee-based methods also offer some promise, but may be limited by difficulties in the consistent identification of sufficient marginal or unsatisfactory performers. Outcome-based methods have not yet been extensively developed for use in setting job performance standards, but also appear promising. All three approaches warrant further investigation.

Several aspects of the judgment process were identified as important in standard setting. These include the number and characteristics of the judges used, the training provided to judges, and the use of judgment facilitation techniques including group discussions or the provision of normative data. Research will be required to identify appropriate judges for setting Army enlisted performance standards and to develop specific procedures for training judges and facilitating their judgments.

Procedures for combining multiple performance standards into an overall standard were also reviewed. Some compromise between a strictly compensatory model and a multiple hurdles model would appear to offer the most promise for use with job performance standards. Research to evaluate a range of alternatives is required.

Finally, procedures for linking job performance standards to selection test standards were reviewed. Evaluation of the costs of alternative types of classification errors are required by all of the models reviewed. The problem is further complicated by the need to assure that each job has an adequate supply of future leaders as well as first-tour performers. Opportunities exist for building upon the Project A utility research, but additional research will also be required.

REFERENCES

Alley, William E. (1987, June). Occupational Learning Difficulty. Paper prepared for the Linking Issues Workshop of the National Academy of Sciences Committee on the Performance of Military Personnel, Sante Fe, NM.

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 45-50.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (pp. 508-600). Washington, DC: American Council on Education.

Arabian, J. (1986). Standard setting: An annotated bibliography (Working Paper No: RS-WP-07-8). U. S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

Bernardin, H. J. (1978). Behavioral expectation versus summated scales: A fairer comparison. Journal of Applied Psychology, 63, 125-131.

Bernardin, H.J., & Walter, C.S. (1977). Effects of rater training and diary keeping on psychometric error in ratings. Journal of Applied Psychology, 62, 64-69.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. Journal of Experimental Education, 45, 4-9.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.

Block, J. H. (1978). Standards and criteria: A response. Journal of Educational Measurement, 15, 291-295.

Borman, W.C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 60, 556-560.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 42, 219-240.

Cangelosi, J. S. (1984). Another answer to the cut-off score question. Educational Measurement: Issues and Practice, 3, 23-25.

Cross, L.H., Impara, J.C., Frary, R.B., & Jaeger, R.M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Journal of Educational Measurement, 21, 113-130.

Dalkey, N. (1969). The Delphi method: An experimental study of group opinions. Santa Monica, CA: Rand Corp.

de Gruijter, D. N., & Hambleton, R. K. (1984). On problems encountered using decision theory to set cutoff scores. Applied Psychological Measurement, 8(1), 1-8.

DiNisi, A.S., Cafferty, T.P., & Meglino, B.M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. Organizational Behavior and Human Performance, 33, 360-369.

Ebel, R. L. (1972). Essentials of Educational Measurement (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Edwards, W. (1971). Social Utilities. The Engineering Economist, Summer Symposium Series, 6.

Feldman, T.M. (1986). Instrumentation and training for performance appraisal: A perceptual-cognitive viewpoint. In K.M. Rowland and J.R. Ferris (Eds), Research in personnel and human resource management (vol. 4). Greenwich, CT: JAI Press.

Gardiner, P.C., & Edwards, W. (1975). Public values: Multi-attribute utility measurement for social behavior. In M.F. Kaplan and S. Schwartz (Eds.), Human judgment and decision processes. New York: Academic Press.

Glass, G. V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-262.

Green, P.E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. Journal of Consumer Research. 5, 103-123.

Goldstein, I.L. (1986). Training in organizations: Needs assessment development, and evaluation. Monterey, CA: Brooks/Cole.

Halpin, G., & Halpin, G. (1983, August). Reliability and validity of 10 different standard setting procedures. Paper presented at the 91st Annual Meeting of the American Psychological Association, Anaheim, CA.

Hambleton, R. K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 15, 277-290.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art (pp. 80-123). Baltimore, MD: Johns Hopkins University Press.

Hambleton, R. K., & Eignor, D. R. (1979). Issues and methods for standard setting. In AERA Training Program Materials, Criterion-referenced test development and validation methods (Unit 6). Unpublished training materials.

Hambleton, R. K., & Eignor, D. R. (1980). Competency test development, validation, and standard setting. In R. M. Jaeger & C. K. Tittle (Eds.), Minimum competency achievement testing: Motives, models, measures, and consequences (pp. 367-396). Berkeley, CA: McCutchan Publishing Corporation.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10(3), 159-170.

Hambleton, R. K., & Powell, S. (1983). A framework for viewing the process of standard setting. Evaluation and the Health Professions, 6(1), 3-24.

Hanser, L. M. & Grafton, F. G. (1983). Predicting job proficiency in the Army: Race, sex and education. U.S. Army Research Institute. (Unpublished)

Huyhn, H. (1976). Statistical consideration of mastery scores. Psychometrica, 41, 65-78.

Jaeger, R. M. (1976). Measurement consequences of selected standard setting models. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Jaeger, R. M. (1978). A proposal for setting a standard on the North Carolina high school competency test. Paper presented before the Annual Meeting of the North Carolina Association for Research in Education, Chapel Hill.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4(4), 461-475.

Jaeger, R. M., & Busch, J. C. (1984). The effects of a delphi modification of the Angoff-Jaeger standard-setting procedure on standards recommended for the National Teacher Examinations. Paper presented at the joint annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, LA. (ERIC Document 246 091)

Jaeger, R. M., & McNulty, S. (1986, July). Procedures for eliciting and using judgments of the value of observed behaviors on military job performance tests. Prepared for the Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education, National Research Council/National Academy of Sciences.

Keeney, R. L. (1972). Utility functions for multi-attributed consequences. Management Science, 18, 276-287.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 17, 167-178.

Kriewall, T. E. (1972). Aspects and applications of criterion-referenced tests. Downers Grove, IL: Institute for Educational Research, April, 1972. ERIC No. ED 063 333, 27 pp.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers ot minimize rating errors in the observation of behavior. Journal of Applied Psychology, 60, 550-555.

Livingston, S. A. (1982). Assumptions of standard setting methods. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Livingston, S. A., & Kastrinos, W. (1982). A study of the reliability of Nedelsky's method for choosing a passing score (Report No. ETS-RR-82-6). Princeton, NJ: Educational Testing Service. (ERIC Document No. ED 218 361)

Livingston, S. A., & Zieky, M. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Livingston, S. A., & Zieky, M. (1983). A comparative study of standard-setting methods (Research Report No. 83-38). Princeton, NJ: Educational Testing Service.

Locke, E.A., Shaw, K.N., Saari, L.M., & Latham, G.P. (1981). Goal setting and task performance. Psychological Bulletin, 90, 125-152.

Luce, R. & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. Journal of Mathematical Psychology, 1, 1-27.

MacPherson, D. (1981). Predicting skill qualification test item difficulty from judgments. In S. F. Bolin (Chair), Panel on skill qualification testing: An evolving system (pp. 1383-1390). Arlington, VA: 23rd Annual Conference of the Military Testing Association. (DTIC, ADP 001400)

McLaughlin, D. H., Rossmeissl, P. G., Wise, L.L., Brandt, D. A. & Wang, M. (1984) Development and validation of Army selection and classification measures. Project A: Validation of current and alternative ASVAB Area Composites, based on training and SQT information on FY81 and FY82 Enlisted accessions. Alexandria, VA: U.S. Army Research Institute Technical Report 651. (AD A156 807)

Mellenbergh, G. J., Koppelaar, H., & Van der Linden, W. T. (1977). Dichotomass decisions based on dichotomously scored items: A case study, Statistica Neerlandica, 31, 161-169.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 46, 133-158.

Meskauskas, J. A. (1983). Standard setting: State of the art, future prospects. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Norcini, J. J., Lipner, R. S., & Langdon, L. O. (1987). A comparison of three variations on a standard-setting method. Journal of Educational Measurement, 24, 56-64.

Nord & White (1987, June) The measurement and application of performance utility: some key issues. Paper prepared for the Linking Issues Workshop of the National Academy of sciences Committee on the Performance of Military Personnel, Sante Fe, NM

Poggio, J. P., Glassnap, D. R., & Eros, D. S. (1981). An empirical investigation of the Angoff, Ebel, and Nedelsky standard-setting methods. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Poggio, J. P. (1984, April). Practical considerations when setting test standards: A look at the process used in Kansas. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document 249 267)

Popham, W. J. (1978). As always, provocative. Journal of Educational Measurement, 15, 297-300.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. Journal of Applied Psychology, 69, 581-588.

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating formats. Organizational Behavior and Human Decision Processes, 38, 76-91.

Raiffa, H. (1968). Decision analysis: Introductory lectures on choices under certainty. Reading, Mass: Addison-Wesley.

Sadacca, R. A., Park, M. V. & White, L. (1986). Weighting performance constructs in composite measures of job performance. Paper presented at the Annual Meeting of the American Psychological Association, Washington, D.C.

Saunders, J. C., & Helsley, T. L. (1987). Setting standards at the objective level: A critique of the Cangelosi Method. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Schneider, B. & Schmitt, N. (1987). Staffing Organizations. Glenview, IL: Scott-Foresman.

Scriven, M. (1978). How to anchor standards. Journal of Educational Measurement, 15, 273-276.

Shepard, L. A. (1980). Standard setting issues and methods. Applied Psychological Measurement, 4, 447-467.

Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 169-198). Baltimore, MD: John Hopkins University Press.

Shikiar, R., & Saari, L. M. (1985). Establishing cut scores for the NRC reactor operator and senior reactor exam (Technical Evaluation Report No. PNL-5131). Seattle, WA: Pacific Northwest Laboratory.

Sigmon, G. L., & Halpin, G. (1984, April). Application of judgmental standard setting procedures to vocational evaluation competency statements by rehabilitation field personnel and educators. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Skakun, E.N., & Kling, S. (1980). Comparability of methods of standards setting. Journal of Educational Measurement, 17, 229-235.

van der Linden, W. J. & Mellenberg, G. (1977)Optimal cutting scores using a linear loss function. Applied Psychological Measurement, 1 (4), 593-599.

Wigdor, A. K. & Green, B. R. (Eds.) (1986). Assessing the performance of military enlisted personnel. Washington D.C: National Academy Press.

Wexley, K. N., & Latham, G. P. (1981). Developing and training human resources in organizations. Glenview: IL: Scott, Foresman, and Co.

Wise, L. L., Campbell, J. P., & Arabian, J. M. (1987). The Army Synthetic Validation Project. Paper presented at the Linking Issues Workshop organized by the National Academy of Sciences Committee on the Performance of Military Personnel, Sante Fe, NM.

Zieky, M. L., & Livingston, S. A. (1977). Manual for setting standards on the Basic Skills Assessment Tests. Princeton, NJ: Educational Testing Service.